

U.S. PATENT APPLICATION

DISTRIBUTING AND BALANCING TRAFFIC FLOW IN A VIRTUAL GATEWAY

INVENTORS: Douglas McLaggan
80 East Craigs Rigg
Edinburgh, Midlothian EH12 8JA
United Kingdom
A Citizen of Britain

Ian Herbert Wilson
1 Westhall Gardens
Edinburgh, Midlothian EH10 4JJ
United Kingdom
A Citizen of Britain

Mark A. Denny
4289 Christian Drive
San Jose, California 95135
A Citizen of Canada

Rick L. Williams
3341 Ventura Circle
Wake Forest, North Carolina 27587
A Citizen of the United States of America

ASSIGNEE: CISCO TECHNOLOGY, INC.
170 WEST TASMAN DRIVE
SAN JOSE, CA 95134

A CALIFORNIA CORPORATION

ENTITY: LARGE

BEYER WEAVER & THOMAS, L.L.P.
P.O. Box 778
Berkeley, CA 94704-0778
Telephone (510) 843-6200

Distributing And Balancing Traffic Flow In A Virtual Gateway

Douglas McLaggan, Ian H. Wilson, Mark A. Denny, Rick L. Williams

BACKGROUND OF THE INVENTION

The present invention relates generally to network systems using redundant or standby devices working together in a redundancy group and load distributing arrangement to provide a virtual router service. More particularly, the present invention relates to methods and apparatus for controlling the distribution of traffic flow across a gateway using multiple gateway devices that are acting as a virtual router.

Local area networks (LANs) are commonly connected with one another through one or more routers so that a host (a PC or other arbitrary LAN entity) on one LAN can communicate with other hosts on different LANs (that is, remote or external networks). Typically, a host is able to communicate directly only with the entities on its local LAN segment. When it needs to send a data packet to an address that it does not recognize as being local, it communicates through a router (or other layer-3 or gateway device) which determines how to direct the packet between the host and the destination address in a remote network. Unfortunately, a router may, for a variety of reasons, become inoperative after a “trigger event” (for example, a power failure, rebooting, scheduled maintenance, etc.). Such potential router failure has led to the development and use of redundant systems, which have more than one gateway device to provide a back up in the event of primary gateway device failure. When a gateway device fails in such a redundancy system, the host communicating through the inoperative gateway device may still remain connected to other LANs by sending packets to and through another gateway device connected to the host’s LAN.

Logically, such a system can resemble Figure 1A. In Figure 1A, a local network 130 uses a single gateway router 110 to forward outbound packets for hosts 122, 124, 126 when those packets are bound for an outside network 150 (for example, the Internet). As seen in Figure 1B, however, the actual physical configuration of a

redundancy group system uses several routers 112, 114, 116, 118 to implement a redundancy group that functions as the single virtual gateway 110 for hosts 122, 124, 126.

Various protocols have been devised to allow a host to choose a router from among a group of routers in a network. Two of these, Routing Information Protocol (or RIP) and ICMP Router Discovery Protocol (IRDP) are examples of protocols that involve dynamic participation by the host. However, because both RIP and IRDP require that the host be dynamically involved in the router selection, performance may be reduced and special host modifications and management may be required.

In a widely used and somewhat simpler approach, the host recognizes only a single “default” router. Hosts (for example, workstations, users and/or data center servers) using the IP protocol utilize this default gateway to exit a local network and access remote networks. Therefore, each host must have prior knowledge of the gateway’s IP address which typically is a router or layer-3 switch IP address. Hosts are either statically configured with the IP address of the default gateway or are assigned the address through a configuration protocol (such as Cisco’s DHCP) upon boot-up. In either case, the host uses the same default gateway IP address for all network traffic destined to exit the local network.

To forward traffic to the default gateway, the host must perform an IP-ARP resolution to learn the data-link Media Access Control (MAC) address of the default gateway. The host sends an ARP inquiry to the IP address of the gateway, requesting the gateway’s MAC address. The default gateway will respond to the host’s ARP request by notifying the host of the gateway’s MAC address. The host needs the default gateway’s MAC address to forward network traffic to the gateway via a data-link layer transfer. When only a single gateway device is used, that device returns its own “burned in” MAC address (BIA MAC address) as the address for the host’s outbound packets.

In this approach, the host is configured to send data packets to the default router when it needs to send packets to addresses outside its own LAN. It does not keep track of available routers or make decisions to switch to different routers. This

requires very little effort on the host's part, but has a serious danger. If the default router fails, the host cannot send packets outside of its LAN. This may be true even though there may be a redundant router able to take over, because the host does not know about the backup. Unfortunately, such systems have been used in mission critical applications.

The shortcomings of these early systems led to the development and implementation of redundant gateway systems, which provide for failover in gateway settings. One such system is the hot standby router protocol (HSRP) by Cisco Systems, Inc. of San Jose, California. A more detailed discussion of the earlier systems and of an HSRP type of system can be found in United States Patent No. 5,473,599 (referred to herein as "the '599 Patent"), entitled STANDBY ROUTER PROTOCOL, issued Dec. 5, 1995 to Cisco Systems, Inc., which is incorporated herein by reference in its entirety for all purposes. Also, HSRP is described in detail in RFC 2281, entitled "Cisco Hot Standby Router Protocol (HSRP)", by T. Li, B. Cole, P. Morton and D. Li, which is incorporated herein by reference in its entirety for all purposes.

HSRP is widely used to back up primary routers for a network segment. In HSRP, a "standby" router is designated as the backup to an "active" router. The standby router is linked to the network segment(s) serviced by the active router. The active and standby routers share a single "virtual IP address" and, possibly, a single "virtual Media Access Control (MAC) address" which is actually in use by only one router at a time. All internet communication from the relevant local network employs the virtual IP address (also referred to as a "vIP address") and the virtual MAC address (also referred to herein as a "vMAC address"). At any given time, the active router is the only router using the virtual address(es). Then, if the active router should cease operation for any reason, the standby router immediately takes over the failed router's load (by adopting the virtual addresses), allowing hosts to always direct data packets to an operational router without monitoring the routers of the network.

One drawback to HSRP systems in general is that only one gateway device in a redundancy group is in use at any given time. To better utilize system resources in such redundancy systems, a gateway load balancing protocol (GLBP) was developed

by Cisco and is the subject of commonly owned and copending United States Serial No. 09/883,674 filed June 18, 2001, entitled GATEWAY LOAD BALANCING PROTOCOL, which is incorporated herein by reference in its entirety for all purposes.

It should be noted here that the term "gateway load balancing protocol" is somewhat of a misnomer (or at least is not as precise as it might be). While the members of a redundancy group share the traffic flow, there has been no "balancing" of the traffic loads, *per se*, across the gateway. It is true that sharing the traffic load among members of a redundancy group means that responsibility for all traffic is not borne by a single gateway device. However, the terms "load sharing" and "load distribution" more accurately describe the actual implementations of these earlier systems. Therefore, the terms "load sharing" and "load distribution" and the like herein mean the ability to assign outgoing traffic to multiple gateway devices so that a single gateway device is not responsible for all outbound packets from all hosts on a LAN. (For the sake of reference to previously filed patent applications and other publications relied upon herein, the acronym GLBP will still be used herein to refer to the earlier, basic underlying load sharing protocol developed by Cisco Systems.)

Like HSRP, for communications directed outside of a LAN, GLBP uses a single vIP address shared by multiple redundancy group gateway devices (for example, routers), which also maintain actual IP addresses as well (also referred to as "aIP addresses"). Each gateway device also has its own BIA (actual) MAC address (also referred to herein as an "aMAC address") and a single virtual MAC address. Use of vMAC addresses allows interchangeability of routers without the need for reprogramming of the system.

Each GLBP system has a "master" gateway device (also referred to herein as an "Active Virtual Gateway" or AVG device) in the redundancy group that controls address assignment (ARP responses) and failover features. The AVG instructs an ARPing host to address outgoing communications to a virtual MAC address assigned to one of the redundancy group gateway devices (gateway devices not functioning as a master device may be referred to as "standby" and/or "slave" gateway devices, in accordance with standard GLBP nomenclature and operation). Any gateway device that is forwarding packets is referred to herein as an "Active Virtual Forwarder" or

AVF device. Each redundancy group therefore has one AVG device and one or more AVF devices.

More specifically, a host sends an ARP message to the redundancy group's virtual IP address when the host wants to send a packet outside the local network. The AVG selects an AVF to handle outgoing packets for the host and sends the host a reply message containing the vMAC of the AVF selected by the AVG. The host populates its ARP cache with this vMAC address. Thereafter, the host addresses its outbound packets to the vMAC address in its ARP cache, thus sending these packets to the assigned AVF/router.

In earlier systems, hosts were assigned vMAC addresses by random assignment, round robin assignment or by using another prescribed algorithm or methodology. In the event that an assigned AVF of the group failed, the outgoing communications that were to be handled by the failed AVF had to be sent elsewhere. Upon failure of the originally assigned AVF, the failed AVF's vMAC address was re-assigned to another router, for example another router that is acting as an AVF. Thereafter, outgoing packets from the host (and any other host(s) which send packets to the re-assigned vMAC address) were routed instead to the new owner of that newly re-assigned vMAC address. In the event that the AVG itself failed, additional steps were taken to appoint or elect a new AVG and ensure continuity in the load distribution function. However, if one or more gateway devices took on an inordinate portion of the traffic load, there was no way to balance this load sharing capability to control distribution (evenly or otherwise) the traffic flow through gateway devices at the gateway.

In view of the foregoing, it would be desirable to provide gateway load balancing services for communications from outside a local network while ensuring that redundant, load sharing gateway services are still available for the local network.

SUMMARY OF THE INVENTION

The present invention provides methods, apparatus, products, techniques and systems for controlling the traffic flow across a gateway using multiple gateway devices that are acting as a virtual router. Gateway devices in a redundancy group share responsibility for outgoing packets from hosts through the distribution of forwarding addresses, such as virtual MAC addresses, to gateway devices to which hosts are directed in response to ARP requests.

One aspect of the present invention is a method of controlling traffic flow in a load-sharing redundancy group that includes a first gateway device and a second gateway device, where the gateway devices are configured to forward packets sent from hosts. One group of forwarding addresses is assigned to the first gateway device and a second group of forwarding addresses are assigned to the second gateway device. The redundancy group distributes forwarding addresses to hosts which in turn use the distributed forwarding addresses to send packets to the redundancy group. The traffic flow for each of the assigned forwarding addresses is measured and may, in some cases, be compared to a target traffic flow, which can be a desired traffic balancing among the redundancy group members. Comparison of the measured traffic flow to the target traffic flow may not be necessary in connection with some target traffic flows.

The traffic flow is then adjusted to close in on the target traffic flow. Adjustment of the traffic flow may be accomplished by changing the existing measured traffic flow (for example, by re-assigning a forwarding address having a certain measured traffic on a gateway device to a different gateway device) or by altering the future distribution of forwarding addresses so that additional traffic is sent to the forwarding addresses having lower measured traffic. The gateway devices can be routers using virtual MAC addresses as forwarding addresses. The redundancy group may also be configured to provide failover services in the event that one of the gateway devices ceases operation. A computer program product having a machine readable medium and program instructions contained in the machine readable

medium, may specify one or more of these methods of controlling traffic flow in a load-sharing redundancy group. Similarly, an apparatus for performing such methods of controlling traffic flow in a load-sharing redundancy group also are disclosed.

Another aspect of the present invention pertains to a primary gateway device configured to control traffic flow in a load-sharing redundancy group having the primary gateway device and a secondary gateway device configured to forward packets sent from hosts. The primary gateway device has one or more processors and a memory in communication with at least one of the processors. A least one of the processors and the memory are configured to assign a first group of forwarding addresses to the primary gateway device and to assign a second plurality of forwarding addresses to the secondary gateway device. The primary gateway device distributes forwarding addresses to hosts which use the distributed forwarding addresses to send outgoing packets to the virtual gateway. The traffic flow for each assigned forwarding address in each gateway device is measured and may be compared to a target traffic flow in some cases. The traffic flow is then adjusted.

Adjustment of the traffic flow may be accomplished by the primary gateway device by changing the existing measured traffic flow (for example, by re-assigning a forwarding address having a certain measured traffic on a gateway device to a different gateway device) or by altering the future distribution of forwarding addresses so that additional traffic is sent to the forwarding addresses having lower measured traffic. The primary and secondary gateway devices can be routers using virtual MAC addresses as forwarding addresses. The redundancy group may also be configured to provide failover services in the event that one of the gateway devices ceases operation.

These and other features and advantages of the present invention will be presented in more detail in the following specification of the invention and the associated figures.

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention may best be understood by reference to the following description taken in conjunction with the accompanying drawings in which:

Figure 1A is a schematic diagram of the logical structure a gateway service for hosts in a LAN.

Figure 1B is a schematic diagram of the physical structure of the gateway service shown in Figure 1A in which multiple gateway devices are used in a redundancy group to provide resiliency.

Figure 2 is a schematic diagram of a virtual gateway in which several gateway devices are available to both direct traffic outside the local network and also to control traffic flow to the members of the redundancy group, using the present invention.

Figure 3 is a schematic diagram showing adjustment of traffic flow in a virtual gateway by re-assignment of a forwarding address from one gateway device to a different gateway device.

Figure 4 is a schematic diagram showing adjustment of traffic flow in a virtual gateway by re-assignment of two forwarding addresses from one gateway device to a different gateway device.

Figure 5 is a schematic diagram showing adjustment of traffic flow in a virtual gateway by distribution of a low traffic flow forwarding address to hosts.

Figure 6 is a diagrammatic representation of a gateway device such as a router in which embodiments of the present invention may be implemented.

DETAILED DESCRIPTION OF THE EMBODIMENTS

1. Definitions

Reference will now be made in detail to the preferred embodiment of the invention. An example of the preferred embodiment utilizing products, protocols, methods, systems and other technology developed, sold and/or used by Cisco Systems is illustrated in the accompanying drawings. While the invention will be described in conjunction with that preferred embodiment, it will be understood that it is not intended to limit the invention to one preferred embodiment or to its implementation solely in connection with Cisco products and systems. On the contrary, the following description is intended to cover alternatives, modifications, and equivalents as may be included within the spirit and scope of the invention as defined by the appended claims. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present invention. The present invention may be practiced without some or all of these specific details. In other instances, well known process operations have not been described in detail in order not to unnecessarily obscure the present invention.

The following terms are used in the instant specification. Their definitions are provided to assist in understanding the preferred embodiments described herein, but do not necessarily limit the scope of the invention.

A “host” is a PC, or other arbitrary network entity residing on a LAN, that periodically communicates with network entities outside the LAN on which the host resides through a router or bridge. The term “user” may be used equivalently in this disclosure.

A “router” is a piece of hardware which operates at the network layer to direct packets between various LANs or WANs of a network. The network layer generally allows pairs of entities in a network to communicate with each other by finding a path through a series of connected nodes. Other terms that may be used in this disclosure include “layer 3 switch”, “layer 3 device” and “gateway device” which are not

necessarily the same as a router, but which may function in the same place and manner as a router. Any and all of these terms are intended to be interpreted as broadly as possible, unless specifically defined more narrowly.

An “IP (internet protocol) address” is a network layer address for a device operating in the IP suite of protocols. The IP address is typically a 32 bit field, at least a portion of which contains information corresponding to its particular network segment. Thus, the IP address of a router may change depending upon its location in a network. An IP address that is referred to as “unique” may be globally unique or may be sufficiently unique for the uses for which it is intended (for example, in a limited network environment in which globally unique IP addresses are unnecessary, but in which local IP addresses used on a local network are not sufficiently unique).

A “MAC address” is a data link layer device address, defined by the IEEE 802 committee that deals with issues specific to a particular type of LAN. The types of LANs for which MAC addresses are available include token ring, FDDI and Ethernet. A MAC address generally is intended to apply to a specific physical device no matter where it is plugged into the network. Thus, a MAC address generally is hardcoded into the device -- on a router’s ROM, for example. This should be distinguished from the case of a network layer address, described above, which changes depending upon where it is plugged into the network. Also, a real MAC address (such as a “burned in address” or BIA MAC address) can be distinguished from a “virtual address” (as defined below) which can include a virtual MAC address.

A “virtual address” is typically (but not necessarily) an address shared or able to be shared (for example, via portability of the address through re-assignment) by a group of real network entities, corresponding to a virtual entity. For example, in the context of this invention, one real router from among two or more real routers emulates a virtual router by adopting a virtual address (such as a virtual IP address), and another entity (usually a host) is configured to send data packets to such virtual address, regardless of which real router is currently emulating the virtual router. In the preferred embodiments, the virtual addresses may encompass both MAC and IP addresses. Usually, various members of the group each have the capability of adopting the virtual address(es) to emulate a virtual entity.

A “packet” is a collection of data and control information including source and destination node addresses, formatted for transmission from one node to another. In the context of this invention, it is important to note that hosts on one LAN send packets to hosts on another LAN through a router or bridge connecting the LANs.

2. Overview

Embodiments of the present invention use gateway devices each having been assigned a set of forwarding addresses each of which includes one or more forwarding addresses (for example, vMAC or other MAC addresses) to control distribution of traffic flow consisting of packets that are sent by hosts across multiple gateway devices that are functioning as one or more virtual gateways for communications outside a local subnet. One embodiment of the present invention uses re-assignment of a vMAC address from one gateway device to another to shift the traffic assigned to that vMAC address from the former gateway device to the latter. Another embodiment of the present invention uses the assignment of vMAC addresses during the Address Resolution Protocol (ARP) process to allocate more traffic to one or more vMAC addresses where higher traffic flow is desired. Other embodiments and variations of the present invention will be apparent to those skilled in the art after reading the present disclosure in connection with the Figures.

More specifically, the present invention can be used to control the distribution (for example, load balancing) of traffic loads among gateway devices that are members of a redundancy group sharing a virtual IP address. Traffic flow data for the vMAC addresses is measured by a collector (for example, Cisco’s Netflow, modified to maintain the number of data packets and bytes received for each virtual MAC address, rather than the group vIP address) which collects traffic flow data for packets received by the redundancy group’s vIP address. This data can be shared among redundancy group gateway devices and adjustments can be made to achieve a target

traffic flow that may include desired load sharing and/or distribution characteristics (for example, maintaining generally even balancing of traffic loads) based on the available traffic data. Those skilled in the art will appreciate that various data collection and data sharing techniques and apparatus can be used and different load sharing and balancing criteria applied to implement the present invention and thus are included within the scope of the present invention.

In one embodiment of the present invention, measured traffic flow data includes the rate of traffic on a per destination (that is, forwarding address, such as a vMAC or other MAC address) basis. Each gateway device may regularly poll its own traffic data collector 250 shown in Figure 2 (or any other suitable traffic data collector, such as a single traffic data collector for all members of the redundancy group) to measure the traffic flow rate for each forwarding address that is assigned to that gateway device. Measured traffic flow data can be appended to GLBP Hello messages, or be sent to other gateway devices or a control gateway device (for example, the AVG) in any other suitable way. Thus, at any given time, at least one redundancy group router (or other gateway device) knows the measured traffic flow (that is, how much traffic is being sent to each forwarding address in the redundancy group).

In an embodiment of the present invention shown in Figure 2, the redundancy group is a GLBP group using routers 212, 216 as gateway devices (though expansion of the application of the present invention to redundancy groups having more than two gateway devices will be clearly apparent to those skilled in the art). Each router 212, 216 in the GLBP group is an AVF and initially is assigned a set of vMAC addresses. The vMAC addresses are used as forwarding addresses by hosts to which those forwarding addresses have been distributed in ARP replies. In prior systems, each AVF/gateway device had only been assigned a single vMAC address. In such prior systems, an AVF's vMAC address would be re-assigned (that is, transferred to a different gateway device) only after a trigger event that stopped the AVF from forwarding packets sent to it.

In the embodiment of the present invention shown in Figure 2, by replying to ARP requests from hosts, the AVG 212 controls distribution of vMAC addresses to

hosts using the virtual gateway implemented by redundancy group 210. Members of the redundancy group can be notified which router in the GLBP group is the AVG for that group using services and/or protocols known to those skilled in the art. When a new gateway device enters service, it can register with the AVG to receive an initial set of vMAC addresses. Other methods can be used to make initial vMAC address assignments, as will be appreciated by those skilled in the art.

As seen in Figure 2, virtual gateway 210 is made up of two actual gateway devices (routers) 212 and 216. Each gateway device has an aIP address, an aMAC address and a set of two or more virtual MAC addresses (also referred to as vMAC addresses) used as forwarding addresses and initially assigned to the gateway device. The gateway devices in redundancy group 210 share vIP address 10.0.0.100. As illustrated, gateway device 212 uses aIP address 10.0.0.254, aMAC address 0000.0C12.3456 and 5 different vMAC addresses ranging from 0007.B400.0101 through 0007.B400.0105; gateway device 216 uses aIP address 10.0.0.252, aMAC address 0000.0CDE.F123 and 5 different vMAC addresses ranging from 0007.B400.0106 through 0007.B400.0110.

The local subnet 230 that virtual router 210 serves includes hosts 222, 224 and 226, which each have an IP address and a MAC address. For example, host 224 has an IP address of 10.0.0.2 and a MAC address of AAAA.AAAA.AA02. As with some prior GLBP systems, the hosts may have been pre-programmed with the gateway address of the virtual router, in this case 10.0.0.100. An ARP resolution protocol and apparatus similar to a standard Cisco GLBP system can be used in connection with the present invention.

As an illustration, when host 224 sends an ARP request, indicated by arrow step 241, to the gateway IP address (for example, 10.0.0.100), only the AVG 212 responds, distributing one of the forwarding addresses currently assigned to gateway device 216 (for example, vMAC address 0007.B400.0108) to the requesting host 224, at step 242. At step 243, the host 224 caches this vMAC address as the MAC address that corresponds to the default gateway IP address, and then, at step 244, sends packets destined for a network outside the LAN to the cached vMAC address, which currently is assigned to gateway device 216.

As seen in Figure 2, using similar ARP techniques, host 222 has cached vMAC address 0007.B400.104 (a vMAC address currently assigned to gateway device 212) and host 226 has cached vMAC address 0007.B400.0102 (a vMAC address also currently assigned to gateway device 212). Additional hosts, gateway devices and/or vMAC addresses can be included in such a configuration, as will be appreciated by those skilled in the art. Increasing the number of vMAC addresses per gateway device in a gateway improves the granularity of the load control of a given gateway, thus affording “finer tuning” of the control over a gateway’s traffic load distribution.

The AVG may reply to ARP requests, distributing vMAC addresses to hosts, in various ways. For example, the AVG may initially use a round robin algorithm, random distribution algorithm or other technique or algorithm to distribute vMAC addresses to hosts as evenly as possible when initially replying to ARPs from hosts. As discussed below in more detail, the manner of allocating traffic to vMAC addresses may change over time or may remain the same. If vMAC addresses are distributed to hosts in ARP replies in a relatively even initial distribution, and each host is sending the same traffic load to its assigned vMAC address, then the traffic loads of the gateway devices and vMAC addresses illustrated in the example of Figure 2 will be the same, as shown in Table 1.

TABLE 1

| Router 212 | | Router 216 | |
|-----------------------|----------------|-----------------------|----------------|
| <u>vMAC addresses</u> | <u>% Loads</u> | <u>vMAC addresses</u> | <u>% Loads</u> |
| 0007.B400.0101 | 10.0 | 0007.B400.0106 | 10.0 |
| 0007.B400.0102 | 10.0 | 0007.B400.0107 | 10.0 |
| 0007.B400.0103 | 10.0 | 0007.B400.0108 | 10.0 |
| 0007.B400.0104 | 10.0 | 0007.B400.0109 | 10.0 |
| 0007.B400.0105 | 10.0 | 0007.B400.0110 | 10.0 |
| Total | 50.0 % | Total | 50.0 % |

However, traffic loads may change over time. For example, some hosts may make heavier use of the gateway, some hosts may cease sending packets to their cached vMAC address(es), etc. As a result, one of the gateway devices (or, more precisely, one or more of the vMAC addresses assigned to the gateway device) in

Figure 2 may eventually be forwarding more traffic than another, one example of which is shown in Table 2.

TABLE 2

| Router 212 | | Router 216 | |
|-----------------------|----------------|-----------------------|----------------|
| <u>vMAC addresses</u> | <u>% Loads</u> | <u>vMAC addresses</u> | <u>% Loads</u> |
| 0007.B400.0101 | 20.0 | 0007.B400.0106 | 10.0 |
| 0007.B400.0102 | 13.0 | 0007.B400.0107 | 5.0 |
| 0007.B400.0103 | 15.0 | 0007.B400.0108 | 0.0 |
| 0007.B400.0104 | 15.0 | 0007.B400.0109 | 10.0 |
| 0007.B400.0105 | 7.0 | 0007.B400.0110 | 5.0 |
| Total | 70.0 % | Total | 30.0 % |

One embodiment of the present invention compares the measured traffic flow to a target traffic flow and then adjusts the traffic flow by re-assigning one or more forwarding addresses from a higher traffic flow volume gateway device to a lower traffic flow volume gateway device to close in on or achieve a target traffic flow to the gateway devices making up the gateway. The comparison of the measured traffic flow and the target traffic flow can be performed in the AVG, an AVF or in a comparator outside the redundancy group. If re-assignment of forwarding addresses is being used to adjust the traffic flow of the redundancy group, then the comparison of the measured traffic flow and target traffic flow can be performed in the traffic flow data collector 250 of a gateway device (which may be, for example, a CPU 662 having a memory and processor, as seen in Figure 6). Likewise, if adjustment of the traffic flow is performed, that can be performed in any of the embodiments of the present invention by the AVG, an AVF or any other suitable device in or outside the redundancy group, as will be appreciated by those skilled in the art. If equal distribution of traffic across all gateway devices in a virtual gateway is the target traffic flow, then the measured traffic flow presented in Table 2 can be altered in several ways.

According to one embodiment of the present invention, as seen in Figure 3, traffic information or data (for example, traffic statistics collected on the GLBP interface that record the rate of traffic on a per destination MAC address basis so that the volume of traffic being sent to each vMAC address can be compared) is

exchanged by the gateway devices 212, 216 by hello messages 310. Based on the data collected by and exchanged between the gateway devices (for example, by data collectors 250 in each gateway device 212, 216), the measured traffic flow can be adjusted by re-assigning address 0007.B400.0101 from the forwarding address set of gateway device 212 to the forwarding address set of gateway device 216 at step 320, yielding the vMAC address assignments and traffic load distribution shown in Table 3 and Figure 3. This re-assignment of traffic to a lower volume gateway device is transparent to and does not affect the traffic being generated by the hosts 222, 224, 226 in Figure 3 nor does re-assignment of the vMAC address affect the vMAC addresses cached in each host. Moreover, the AVG does not have to change its assignment algorithm or methodology since the traffic flow will be adjusted by re-assignment to compensate for future imbalance(s), if necessary.

TABLE 3

| Router 212 | | Router 216 | | |
|---------------------|---------------|----------------------|----------------|------|
| <u>vMAC address</u> | <u>% Load</u> | <u>vMAC address</u> | <u>% Load</u> | |
| 0007.B400.0101 | 20.0 | → → | 0007.B400.0106 | 10.0 |
| 0007.B400.0102 | 13.0 | ↓ | 0007.B400.0107 | 5.0 |
| 0007.B400.0103 | 15.0 | ↓ | 0007.B400.0108 | 0.0 |
| 0007.B400.0104 | 15.0 | ↓ | 0007.B400.0109 | 10.0 |
| 0007.B400.0105 | 7.0 | ↓ | 0007.B400.0110 | 5.0 |
| | | → → → 0007.B400.0101 | 20.0 | |
| Total | 50.0 % | Total | 50.0 % | |

Another embodiment of the present invention using vMAC address transfer to adjust traffic flow permits more flexibility with regard to future balancing. After exchanging traffic data at 410, as seen in Figure 4, more than one vMAC address might be re-assigned at 420, for example vMAC addresses 0007.B400.0102 and 0007.B400.0105, thereby shifting 20% of the load to gateway device 216, but allowing for “finer tuning” of the load at a later time, should partial shifting back to gateway device 212 be desired. The resulting vMAC address assignments and measured traffic flow are shown in Table 4 and Figure 4. In this case, shifting traffic to a lower volume gateway device means that packets sent by host 226 at step 430 will

now be forwarded by gateway device 216, instead of gateway device 212, which previously had been forwarding the packets from host 226.

TABLE 4

| Router 212 | | Router 216 | |
|---------------------|---------------|---------------------|----------------|
| <u>vMAC address</u> | <u>% Load</u> | <u>vMAC address</u> | <u>% Load</u> |
| 0007.B400.0101 | 20.0 | 0007.B400.0106 | 10.0 |
| 0007.B400.0102 | 13.0 | → → → | 0007.B400.0107 |
| 0007.B400.0103 | 15.0 | ↓ | 0007.B400.0108 |
| 0007.B400.0104 | 15.0 | ↓ | 0007.B400.0109 |
| 0007.B400.0105 | 7.0 | → | 0007.B400.0110 |
| | | ↓ | 0007.B400.0102 |
| | | → → → | 13.0 |
| | | → → → | 0007.B400.0105 |
| Total | 50.0 % | Total | 50.0 % |

In another embodiment of the present invention, the system can adjust the traffic flow by changing how forwarding addresses are distributed to hosts, rather than re-assigning one or more forwarding addresses to a different gateway device. In the examples shown in Tables 1 and 2 above, for example, the vMAC addresses in use thus remain assigned to the forwarding address sets in the gateway devices to which they were assigned originally. However, the ARP reply algorithm generates replies containing one or more forwarding addresses in the gateway devices to adjust the traffic flow. In other words, regardless of what algorithm or methodology might have been used for initial assignment of hosts to vMAC addresses, the ARP replies can be used to adjust the measured traffic flow by changing the destinations of future packets from one or more hosts.

A generally equal traffic flow from the initial distribution of forwarding addresses to hosts can change as discussed in connection with Tables 1 and 2, above. Using the statistical data shown in Table 2, for example, and the method illustrated in Figure 5, the traffic flow can be adjusted to correct a disparity between the measured traffic flow and the target traffic flow (here, equal distribution of traffic across both gateway devices 212, 216). The method starts at 510 with an existing measured traffic flow. Whenever a host ARPs at 520, the lowest traffic flow gateway device is found first at 530. The lowest traffic flow forwarding address is then found on that gateway

device. The AVG responds to the host's ARP inquiry at 540 by replying with this "lowest percentage" forwarding address as the forwarding address to which the host should send outgoing packets.

In the case shown in Table 2, by responding to ARP requests with the vMAC address having the lowest traffic flow in the gateway device having the lowest traffic flow, in this case the AVG 212 would reply with vMAC address 0007.B400.0108 in AVF 216. The AVG 212 can continue using this forwarding address distribution algorithm or methodology until the measured traffic flows on gateway device 212 and gateway device 216 are within a prescribed percentage or threshold (for example, until the traffic load on each gateway device is at least 45% of the total traffic load). If new traffic is directed to vMAC address 0007.B400.0108, then measured traffic loads on one or more of the other vMAC addresses will decrease at least slightly as the newly distributed address gets more host traffic. The AVG 212 can continue adjusting the traffic flow by making such ARP assignments to close in on or maintain the target traffic flow.

As will be appreciated by those skilled in the art, this "low percentage" assignment method can be used from the outset of operation of the redundancy group 210 to build and maintain a generally equal distribution of traffic flow across a group of gateway devices. If the AVG always distributes the lowest traffic flow forwarding address in the lowest traffic flow gateway device, the system will be self-adjusting to some degree since it will be comparing the measured traffic flow with the target traffic flow whenever a host ARPs for a forwarding address and will adjust the traffic flow merely by distributing the "low percentage" forwarding address. If the measured traffic flow varies too greatly from the target traffic flow, then adjustment of the traffic flow by re-assignment of one or more vMAC addresses, as discussed above, can be used to correct a gross disparity. This hybrid of low percentage address distribution and corrective re-assignment is both simple and effective in establishing, maintaining and adjusting traffic flow to control that traffic flow in a redundancy group.

For example, the AVG may only assign new traffic to a vMAC address until the traffic load on that vMAC address surpasses the next lowest load on the lower

volume gateway device. In Figure 2, if the traffic load on vMAC address 0007.B400.0108 increases to a level above 5.0%, then the AVG may start replying to ARP requests with vMAC address 0007.B400.0107 or 0007.B400.0110, assuming their traffic flows are lower than that of 0007.B400.0108 and that the traffic flow across AVF 216 is still lower than AVG 212. This technique not only balances traffic loads between gateway devices 212 and 216, but also helps to balance the loads among the various vMAC addresses, if that is desired.

In embodiments where variances from the target traffic flow are addressed by distributing low traffic flow forwarding addresses, finer adjustment of the traffic flow can be achieved through the use of more forwarding addresses in each gateway device's forwarding address set. As will apparent to those skilled in the art, each gateway device can conceivably operate with only one forwarding address, provided that hosts regularly send ARP requests for the vIP address. If this is not the case, then adjustment of the traffic flow using re-assignment of forwarding addresses is more effective since correction can be made proactively as soon as an imbalance is detected rather than waiting for an ARP request from a host to begin traffic flow adjustment.

Various types of forwarding addresses will be apparent to those skilled in the art. vMAC addresses are desirable even in the simplest implementation due to the ease with which gateway devices can be replaced and/or added without the need for reprogramming the entire system with new MAC address information. A new gateway device is merely assigned a new set of forwarding addresses comprising one or more vMAC addresses at the time the gateway device is installed into the redundancy group. The use of multiple vMAC addresses in the redundancy group allows easy expansion of the number of gateway devices in the redundancy group and adjustments to operation of the group as well. The use of multiple vMAC addresses within each gateway device provides finer tuning and adjustment of the traffic distribution over the individual gateway devices without adding substantially to the complexity of the system.

Those skilled in the art will appreciate that the selection of a redundancy protocol for use in a given setting may be influenced by various considerations. Regardless of the protocol(s) selected, however, the present invention can be

implemented in a virtual router group or other redundancy group by appropriate selection of the redundancy protocol and the requirements of the system so that the desired performance can be obtained.

Generally, the techniques for implementing the present invention may be implemented on software and/or hardware. For example, these techniques can be implemented in an operating system kernel, in a separate user process, in a library package bound into network applications, on a specially constructed machine, or on a network interface card. In a specific embodiment of this invention, the techniques of the present invention are implemented in software such as an operating system or in an application running on an operating system.

A software or software/hardware hybrid packet processing system of this invention is preferably implemented on a general-purpose programmable machine selectively activated or reconfigured by a computer program stored in memory. Such programmable machine may be a network device designed to handle network traffic. Such network devices typically have multiple network interfaces including frame relay and ISDN interfaces, for example. Specific examples of such network devices include routers and switches. For example, the packet processing systems of this invention may be specially configured routers such as specially configured router models 1600, 2500, 2600, 3600, 4500, 4700, 7200, 7500, and 12000 available from Cisco Systems, Inc. of San Jose, California. A general architecture for some of these machines will appear from the description given below. In an alternative embodiment, the system may be implemented on a general-purpose network host machine such as a personal computer or workstation. Further, the invention may be at least partially implemented on a card (for example, an interface card) for a network device or a general-purpose computing device.

Referring now to Figure 6, a router 610 suitable for implementing the present invention includes a master central processing unit (CPU) 662, interfaces 668, and a bus 615 (for example, a PCI bus). When acting under the control of appropriate software or firmware, the CPU 662 is responsible for such router tasks as routing table computations and network management. It may also be responsible for network address translation, virtual gateway operations, etc. It preferably accomplishes all

these functions under the control of software including an operating system (e.g., the Internet Operating System (IOS.RTM.) of Cisco Systems, Inc.) and any appropriate applications software. CPU 662 may include one or more processors 663 such as a processor from the Motorola family of microprocessors or the MIPS family of microprocessors. In an alternative embodiment, processor 663 is specially designed hardware for controlling the operations of router 610. In a preferred embodiment, a memory 661 (such as non-volatile RAM and/or ROM) also forms part of CPU 662. However, there are many different ways in which memory could be coupled to the system.

The interfaces 668 are typically provided as interface cards (sometimes referred to as "line cards"). Generally, they control the sending and receiving of data packets over the network and sometimes support other peripherals used with the router 610. Among the interfaces that may be provided are Ethernet interfaces, frame relay interfaces, cable interfaces, DSL interfaces, token ring interfaces, and the like. In addition, various very high-speed interfaces may be provided such as fast Ethernet interfaces, Gigabit Ethernet interfaces, ATM interfaces, HSSI interfaces, POS interfaces, FDDI interfaces and the like. Generally, these interfaces may include ports appropriate for communication with the appropriate media. In some cases, they may also include an independent processor and, in some instances, volatile RAM. The independent processors may control such communications intensive tasks as packet switching, media control and management. By providing separate processors for the communications intensive tasks, these interfaces allow the master microprocessor 662 to efficiently perform routing computations, network diagnostics, security functions, etc.

Although the system shown in Figure 6 is one preferred router of the present invention, it is by no means the only gateway device or router architecture on which the present invention can be implemented. For example, an architecture having a single processor that handles communications as well as routing computations, etc. can be used. Further, other types of interfaces and media can also be used with the router.

Regardless of the network device's configuration, it may employ one or more memories or memory modules (including memory 661) configured to store program instructions for the general-purpose network operations and address translation operations described herein. The program instructions may control the operation of an operating system and/or one or more applications, for example. The memory or memories may also be configured to store relevant state information, data structures, etc., such as the measured and target traffic flow data and forwarding addresses described herein.

Because such information and program instructions may be employed to implement the systems/methods described herein, the present invention relates to machine readable media that include program instructions, state information, etc. for performing various operations described herein. Examples of machine-readable media include, but are not limited to, magnetic media such as hard disks, floppy disks, and magnetic tape; optical media such as CD-ROM disks; magneto-optical media such as optical disks; and hardware devices that are specially configured to store and perform program instructions, such as read-only memory devices (ROM) and random access memory (RAM). The invention may also be embodied in a carrier wave traveling over an appropriate medium such as airwaves, optical lines, electric lines, etc. Examples of program instructions include both machine code, such as produced by a compiler, and files containing higher level code that may be executed by the computer using an interpreter.

Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims. Therefore, the described embodiments should be taken as illustrative and not restrictive, and the invention should not be limited to the details given herein but should be defined by the following claims and their full scope of equivalents, whether foreseeable or unforeseeable now or in the future.